***Каніщева О.В.*[1], *Главчева Ю.М.*[1], *Висоцька В.А.*[2]**
[1] *Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;*
[2] *Національний університет "Львівська політехніка", Львів, Україна*

Визначення стилю автора для виявлення плагіату в академічному середовищі

Актуальність роботи. Стратегічне завдання наукових та освітніх закладів – формування нової наукової еліти, яка спроможна внести вклад у розвиток держави шляхом упровадження інноваційних передових розробок у промислове виробництво, сільське господарство, медицину, IT сектор тощо. Ці заходи сприятимуть сталому розвитку економіки держави.

Більшість урядових та громадських організацій спрямовують свої зусилля на розгляд питань, вирішення яких сприятиме створенню умов для підвищення ефективності науки та освіти. Явище плагіату є тим фактором, що негативно впливає на якість науки та освіти [1, 2]. Академічна доброчесність (подібностей) у наукових роботах має сьогодні широке поняття. Це сукупність етичних принципів та визначених законом правил, якими мають керуватися учасники освітнього процесу під час навчання, викладання та провадження наукової (творчої) діяльності з метою забезпечення довіри до результатів навчання та/або наукових (творчих) досягнень. Забезпечення академічної доброчесності вимагає системного підходу до вирішення та реалізації комплексу організаційних, навчальних, технічно-технологічних заходів. Серед останніх важливу роль відіграють: створення депозитарію академічних текстів установи та Національного репозитарію академічних текстів (НРАК); використання інформаційних систем (ІС) виявлення плагіату в наукових роботах, яка буде виявляти подібності на основі даних НРАК та Інтернет.

У роботі [3] плагіат поділяють на чотири види, де кожен має своє цільове призначення. В залежності від типу діяльності та сфери застосування: професійний (присвоєння інтелектуальних, творчих, професійних здобутків інших у професійних цілях), освітньо-науковий (присвоєння чужого інтелектуального майна виключно у процесі здобуття наукового ступеня, освітнього кваліфікації або визнання у цих напрямках), соціальний (аналогічно професійному, але у побутових відносинах і не відноситься до фахової діяльності); нормативний (привласнення законодавчих, юридичних, методичних, наукових, практичних напрацювань).

В залежності від форми авторами цієї роботи виділено наступні види плагіату:

1. Точне/часткове копіювання авторської роботи у мовній, лексичній, технологічній інтерпретації (ідентифікують більшість антиплагіатних ІС).
2. Повторення основної ідеї (визначити доволі складно у зв'язку із відсутністю методів видобування та аналізу змісту документу антиплагіатними ІС).
3. Плагіат, який передбачає наявність певних посилок: у посиланнях; у визначені цитат; посилання на неіснуючі джерела; у наведенні точних фактів з не власних досліджень без конкретизації джерела; у поданих джерелах (частково реалізовано у деяких антиплагіатних ІС).
4. Плагіат як результат замовлення деяким псевдоавтором у комерційній організації або у приватної особи та видачі її за власну роботу (визначення стилю автора, на основі його вже існуючих робіт, що є актуальним завданням і досліджується в межах даної роботи).

Метою даної роботи є визначення стилю автора наукових статей для визначення плагіату на більш глибокому рівні. Цю задачу розглянемо як задачу класифікації текстів (рис. 1), тобто кожний документ відноситься тільки до конкретної категорії/класу (автора). У роботі у якості даних використано одноосібні наукові статті 100 науковців України, з кожного по 10 статей. Таким чином отримано колекцію з 1000 текстів. Для кожної категорії визначено еталонні показники, які характеризують його стиль. Для кожного з досліджених текстів визначено коефіцієнт належності до конкретної категорії. Для визначення стилю автора використані методи машинного навчання (Байєсівський класифікатор та метод опорних векторів). Також досліджено за допомогою нашого корпусу кількісні оцінки мови, такі як

коефіцієнт різноманітності тексту, ступінь синтаксичної складності, зв'язність мовлення, індекс винятковості і концентрації тексту [4], які використовують для визначення стилю автора.

Рис. 1. Загальна схема роботи для визначення стилю автора

Етап предобробки тексту досить трудомісткий та складається з первинного опрацювання лінгвістичних даних (побудова рядів розподілу, обчислення статистик, статистичних оцінок та інші параметри лінгвометрії), лексикографічного опрацювання текстових даних (створення частотних і алфавітно-частотних словників конкретного автора, словників-конкодансів, слововказивників, зворотних словників, словників ключових слів стилю автора тощо) та методів стилеметрії для визначення авторської атрибуції тексту. Без статистично опрацьованого доробку автора (еталону) це практично зробити неможливо. Аналіз та інтерпретація на лінгвістичному рівнів стилістичних особливостей і закономірностей письменницького стилю конкретного автора побудований на методі контент-аналізу та складається з наступних етапів: коректний відбір текстів; лематизація текстових одиниць; усунення неоднорідності текстових одиниць; побудова словників та організація їх на основі статистичних розподілів у потрібних у необхідних частотних словникових шкалах; пошук параметрів, що адекватно відображають структуру частотного словника; перевірка параметрів на ефективність; математичне моделювання та аналіз лексико статистичних розподілів; побудова статистичних класифікацій; інтерпретація результатів та визначення еталонних коефіцієнтів різноманітності тексту конкретного автора. Важливим фактором є побудова коректних словників ключових слів стилю автора. Для досягнення мети дослідження було розроблено систему, що дозволяє обрати мову або мови, з яких складений текст. Доступ до процесу знаходження множини ключових слів з врахуванням основ тематичних слів можна отримати на ресурсі Victana. При побудові будь-яких словників з україномовних текстів необхідно враховувати основи цих слів без флекцій.

Висновки та основні результати. У роботі розглянуто актуальність задачі визначення плагіату, проаналізовано види плагіату, описано підхід для визначення стилю автора. Авторами досліджені методи машинного навчання та кількісні оцінки мови для вирішення задачі визначення стилю автора на прикладі розроблено корпусу українських академічних робіт. Окремою проблемою є не одноосібні публікації, а авторського колективу. Визначення стилю в такому випадку є трудомістким, так як один стиль автора накладається на інший стиль. В такому випадку визначити, що така робота є «комерційною» (замовленою) досить складно.

Література. **1.** Лупаренко Л.А. Інструментарій виявлення плагіату в наукових роботах: аналіз програмних рішень // Інформаційні технології і засоби навчання. – 2014. – Т. 2. – №. 40. – С. 151–169. **2.** Болілий В.О. Перевірка унікальності тексту при оцінюванні студентських робіт творчого або дослідницького характеру / В.О. Болілий, В.В. Копотій // Наукові записки НДУ ім. Гоголя, Психолого-педогогічні науки // Збірник наукових праць. – 2011. – № 7. – С. 134–145. **3.** Петренко В.С. Поняття та види плагіату. / В.С. Петренко // Часопис цивілістики. – 2013. – Вип. 14. – С. 128–131. **4.** Математична лінгвістика. Кн. 1. Квантитативна лінгвістика: навчальний посібник для вузів / За ред., передмова Володимир Володимирович Пасічник; Юрій Миколайович Щербина, Вікторія Анатоліївна Висоцька, Тетяна Валеріївна Шестакевич. – Львів: Новий Світ–2000, 2012. – 358 с.